Claims

1. A method of processing a data record for finding a counterpart in a reference data set, the method comprising the steps of:

determining in the data record a value of a data field, the data field representing an identifier,

determining from a set of predetermined identifier values at least one synonym candidate for the value of the data field,

determining if a synonym candidate and the value of the data field fulfill a predetermined synonym acceptance criterion, and if the predetermined synonym acceptance criterion is fulfilled, associating the value of the data field and the synonym candidate as synonyms, and

searching for a counterpart for the data record by comparing to entries of the reference data set the value of the data field and/or a synonym associated with the value of the data field.

2. A method as defined in claim 1, wherein the at least one synonym candidate is determined using a candidate selection criterion depending at least on the value of the data field and on a synonym candidate.

3. A method as defined in claim 2, wherein the candidate selection criterion takes into account how similar a synonym candidate and the value of the data field sound.

4. A method as defined in claim 2, wherein the candidate selection criterion specifies that at least a predetermined part of the value of the data field is identical to a predetermined part of a synonym candidate.

5. A method as defined in any one of claims 2 to 4, wherein the candidate selection criterion takes into account also a further data field of the data record, said further data field representing a second identifier.

6. A method as defined in any preceding claim, wherein at least one quality parameter is evaluated for a synonym candidate, the synonym acceptance criterion taking into account the at least one quality parameter.

7. A method as defined in claim 6, wherein at least one quality parameter takes into account at least one of the following quantities:
a number of changes required for converting the value of the data field to be identical to a synonym candidate; a proportion of identical characters in the value of the data field and in a synonym candidate; and a difference between the length of the value of the data field and the length of a synonym candidate.

8. A method as defined in claim 7, wherein the number of changes required for converting the value of the data field to be identical to a synonym candidate is calculated using the Levenshtein distance.

9. A method as defined in claim 7, wherein the proportion of identical characters takes into account the order of the characters.

10. A method as defined in any one of claims 6 to 9, wherein a first quality parameter is evaluated for each synonym candidate and at least a second quality parameter is evaluated at least for the synonym candidate(s) having the best first quality parameter.

11. A method as defined in any one of claims 6 to 10, wherein the synonym acceptance criterion requires that there is only one synonym candidate having the best at least one quality parameter.

12. A method as defined in any one of claims 6 to 11, wherein at least two quality parameters are evaluated for each synonym candidate and the synonym candidate acceptance criterion specifies a threshold for one of the at least two quality parameters, the threshold being dependent on a further one of the at least two quality parameters.

13. A method as defined in any preceding claim, wherein the search for the counterpart involves comparison of the value of the data field to a synonym set relating to the

identifier, members of said synonym set referring to respective predetermined identifier values, and when the predetermined synonym acceptance criterion is fulfilled, the value of the data field is added to the synonym set as a member referring to the synonym associated with the value of the data field before the search for the counterpart.

14. A method as defined in any preceding claim, wherein determining the at least one synonym candidate is discarded, if a predetermined discard criterion is fulfilled.

15. A method as defined in claim 14, wherein the predetermined discard criterion specifies that the value of the data field is identical to one of the predetermined identifier values.

16. A method as defined in claim 14, wherein the search for the counterpart involves the synonym set and the predetermined discard criterion specifies that the value of the data field is at least one of the following: one of the predetermined identifier values, and a member of the synonym set.

17. A method as defined in any one of claims 14 to 16, wherein the predetermined discard criterion takes into account a value of a second data field in the data record.

18. A method as defined in any preceding claim, wherein information indicating the at least one synonym associated with the value of the data field is added to the data record.

19. A method as defined in claim 18, wherein a copy of the data record is made for each synonym associated with the value of the data field.

20. A method as defined in any preceding claim, wherein the identifier relates to a name of one of the following: a geographical entity, a person and an organisation.

21. A method of processing a synonym set for searching counterparts in a reference data set for data records, a data record containing a data field representing an identifier, members of the synonym set being first identifier values and referring to respective second identifier values, the second identifier values being predetermined identifier values, and said searching for a counterpart involving comparison of a value of the data field to the

synonym set, the method comprising the steps of determining among the predetermined identifier values at least one synonym candidate relating to the value of the data field in the data record, and, if the value of the data field and a synonym candidate fulfill a predetermined synonym acceptance criterion, adding before searching a counterpart for a data record the value of the data field to the synonym set as a member referring to the synonym candidate.

22. A method as defined in claim 21, wherein the synonym set is empty before adding the value of the data field to the synonym set.

23. A method as defined in claim 21, wherein the synonym set contains at least one member before adding the value of the data field to the synonym set.

24. A computer program comprising program instructions for causing a computer to perform the method of any one of claims 1 to 23.

25. A computer program as defined in claim 24, embodied on a computer-readable record medium.

26. A data processing system for processing data records for finding counterparts in a reference data set, the system comprising:

- means for receiving data records,
- means for storing the reference data set,
- means for storing predetermined identifier values for an identifier,
- means for determining in the data records values of a data field, the data field representing the identifier,
- means for associating values of the data field and respective predetermined identifier values as synonyms, said means configured to determine from the predetermined identifier values at least one synonym candidate for a value of the data field, to determine if a synonym candidate and the value of the data field fulfill a predetermined synonym acceptance criterion, and if the predetermined synonym acceptance criterion is fulfilled, to associate the value of the data field and the synonym candidate as synonyms, and

- means for searching counterparts in the reference data set for the data records, said searching involving comparing to entries of the reference data set values of data fields and/or synonyms associated with the values of the data fields.

27. A data processing system as defined in claim26, further comprising

- means for storing a synonym set, members of said synonym set referring to respective predetermined identifier values,

wherein the means for associating values of the data field and respective predetermined identifier values as synonyms are configured to add to the synonym set a member referring to the synonym associated with the value of the data field before activation of the means for searching counterparts.

28. A data processing system for processing a synonym set for searching counterparts in a reference data set for data records, a data record comprising a data field representing an identifier, members of the synonym set being first identifier values and referring to respective second identifier values, said second identifier values being predetermined identifier values, and said searching involving comparing a value of the data field to the synonym set, the system comprising:

- means for storing the synonym set,
- means for storing predetermined identifier values for the identifier,
- means for receiving data records,
- means for determining in the data records values of the data field, and
- means for adding to the synonym set a value of the data field and respective predetermined identifier values associated as synonyms before searching counterparts in the reference data set, said means configured to determine from the predetermined identifier values at least one synonym candidate for a value of the data field, to determine if a synonym candidate and the value of the data field fulfill a predetermined synonym acceptance criterion, and if the predetermined synonym acceptance criterion is fulfilled, to associate the value of the data field and the synonym candidate as synonyms.